Invariance through Mutual Information Regularization

Justin Tan, Phillip Urquijo

University of Melbourne

April 16, 2019

Flavor Physics

Precision flavor physics

Compare precise experimental measurements of observables in B decays with theoretical predictions; interpret discrepancies in terms of new physics.

- Look for indirect effects of heavy unknown particles in low energy observables of B mesons.

Penguin processes:

Radiative: $b \to q\gamma$

Electroweak:
$b \to q\ell^+\ell^-, \quad q = s, d$

- FCNCs, forbidden at leading order $\to$ rare + hard to observe!

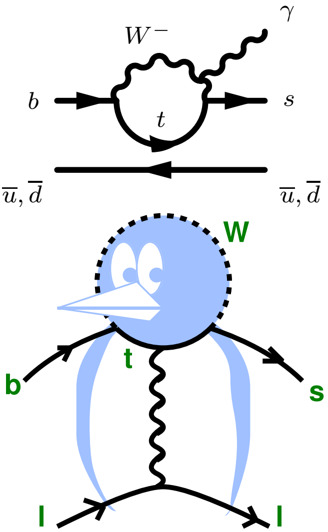


Figure 1: Radiative $b \to s\gamma$ (top) and electroweak $b \to s\ell^+\ell^-$ (bottom) penguins

Belle II

- Next generation B-physics experiment at SuperKEKB, an e^+e^- collider in Japan.
- Target: 50×10^9 $e^+e^- \to \Upsilon(4S) \to B\bar{B}$ events by 2024.
- Large statistics $\to$ high precision measurements of penguin decay observables: $\mathcal{B}(b \to s\gamma), \mathcal{B}(b \to s\ell\ell), R_{X_s}$.

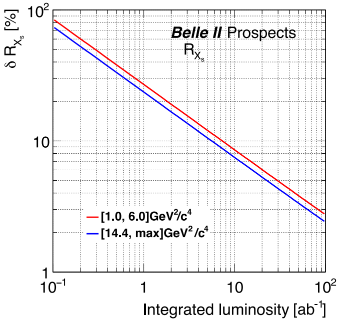


Figure 2: Sensitivity to lepton universality ratio R_{X_s} in different q^2 regions

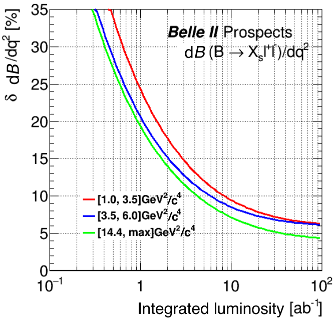
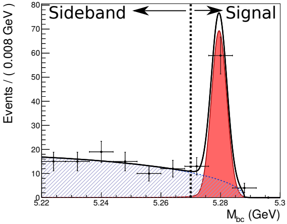


Figure 3: Sensitivity to $d\mathcal{B}/dq^2$ in $b \to s\ell\ell$ decays in different q^2 regions

Signal Identification

- Identify signal peak in:
  - $M_{bc} = \sqrt{E_{beam}^2 - |\vec{p}_B|^2}$
- Extract physical observables by fitting signal + background model.
- Rely on interpolation of smooth background spectrum from sidebands beneath signal peak.



Learning algorithms preferentially select signal-like events $\rightarrow$ background spectrum distortion $\rightarrow$ uncontrollable systematic uncertainties. Necessary to avoid introduction of parameter-dependent bias in signal/background spectrum.

e.g. $b \rightarrow s\ell\ell$ analyses report results in regions of the $q^2 = M_{\ell\ell}^2$ spectrum $\rightarrow$ important that M_{bc} and q^2 should remain unbiased

Setup

- Train supervised learning algorithm to distinguish true signal $b \to s\gamma$ events from background processes.
- Input data X consists of kinematic quantities and event topology variables, ~ 80 in total.
- Sensitive variable Z is the beam constrained mass M_{bc}.
- All variables with Pearson correlation with Z above 0.1 removed.
- Let parameters of the learning algorithm be θ (in this case, a neural network).
- Treat network as an encoder $X \to E$. After training, threshold output E to reject 99.5% of background events.

Sculpting

Classifier output $E_\theta(X) \sim \text{p(signal}|\text{data)}$ (calibration issues aside). Reject given fraction of events by thresholding this output.

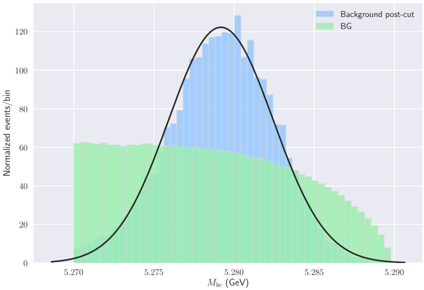


Figure 4: Continuum M_{bc} before (green) and after (blue) @ 0.995 suppression.

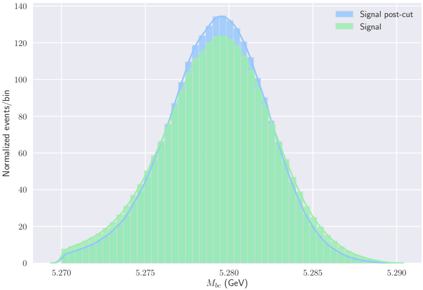


Figure 5: Signal M_{bc} before (green) and after (blue) @ 0.995 suppression.

Background artificially sculpted to resemble signal spectrum post-selection, a result of a non-uniform selection efficiency.

(Non-) Uniform Selection Efficiency

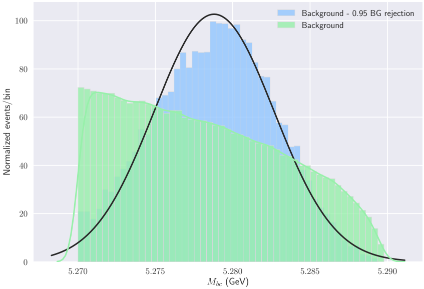


Figure 6: Non-uniform selection efficiency of background events in M_{bc} spectrum.

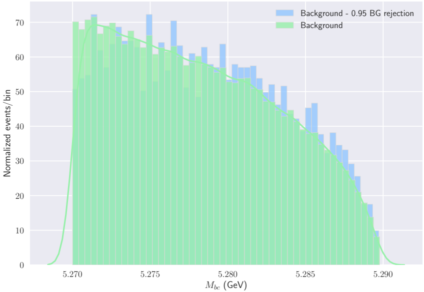


Figure 7: Uniform selection efficiency of background events in M_{bc} spectrum.

- Decay observables measured by conducting a likelihood fit to certain discriminating variables (here M_{bc}).
- Non-uniform selection efficiency in these variables may result in poorly understood systematic uncertainties and increased reliance on (potentially inaccurate) simulated data.

Mutual Information

- Symmetric information measure between random variables X and Y.

$$I(X, Y) \triangleq \mathbb{E}_{X,Y}\left[\log \frac{p(x,y)}{p(x)p(y)}\right]$$
$$= H(Y) - H(Y|X)$$

- $I(X, Y) \geq 0$ with equality if X, Y independent.
- Entropy H is a measure of uncertainty in X:

$$H(X) = -\mathbb{E}_X\left[\log p(x)\right]$$

"Reduction in uncertainty in Y due to knowledge of X."

Mutual Information Penalty

- Treat neural network as an encoder encoding input data $X \to E$.
- Strip dependency of encoding from sensitive variables where uniform selection efficiency desirable (call it Z).
- Augment cross-entropy objective with mutual information between encoder output and variables where uniform selection efficiency is to be enforced.

$$\mathcal{L}\left(\theta_f; Z\right) = H_{p,q} + \lambda I(E, Z)$$

  - $H_{p,q}$: Generic classification loss
  - $I(E, Z)$: Mutual information between encoding E and Z
- Penalize large information content between encoding and Z, penalty strength determined by λ.
- Problem: mutual information intractable to compute in general*

Estimating Mutual Information

- Variational lower bound on the mutual information (Nowozin et. al., NIPS 2016):

$$
\begin{aligned}
I_V(E_{\theta_f}(X), Z) &= \mathbf{E}_{\mathbb{P}_{XZ}}\left[T_\omega(E_\theta(x), z)\right] - \log \mathbf{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}\left[e^{T_\omega(E_\theta(x), z)}\right] \\
&\leq I(E_{\theta_f}(X), Z)
\end{aligned}
$$

- $T_\omega : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$: differentiable transformation parameterized by ω, adjusted to maximize I_V.
- Parameterize mappings E_{θ_f} and T_ω by neural networks.

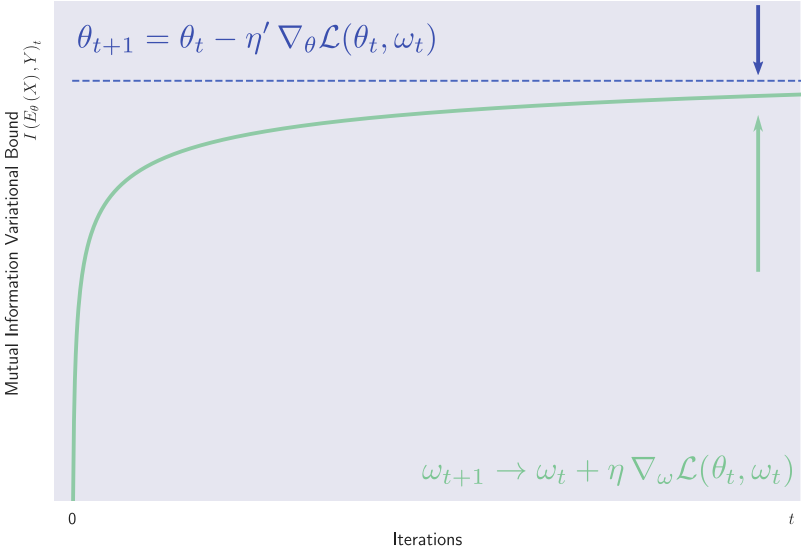
$$
\begin{aligned}
\theta_{t+1} &\leftarrow \theta_t - \eta' \nabla_\theta \mathcal{L}(\theta_t, \omega_t) \\
\omega_{t+1} &\leftarrow \omega_t + \eta \nabla_\omega \mathcal{L}(\theta_t, \omega_t)
\end{aligned}
$$

Mutual Information Penalty

$$\text{Objective: } \min_E \max_T \mathbf{E}_{\mathcal{D}} \left[-\log p_\theta(y|x) \right] + \lambda_{MI} I_V \left(E_{\theta_f}(X), Z \right) \qquad (1)$$

- The classifier/encoder E_{θ_f} enforces decorrelation by minimizing $I(E, Z)$ simultaneously with the cross entropy.
- T_ω tightens the lower bound by maximizing lower bound I_V.
- λ_{MI} controls tradeoff between decorrelation and classification.

Mutual Information Penalty



Toy Example

- Data drawn from bivariate Gaussians:

$$\mathbf{s}_1 \sim \mathcal{N}\left(\begin{pmatrix} z \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right), \quad \mathbf{s}_2 \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 1.5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

- Classify individual samples according to (x, y) coordinates, penalizing dependency on noisy x dimension.

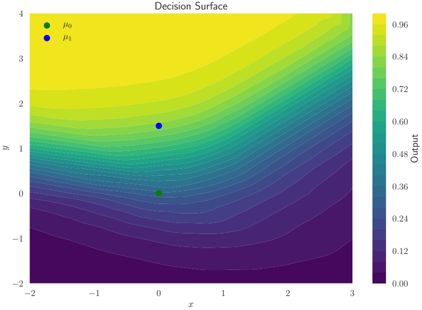


Figure 8: Curvature of contours in decision surface indicates dependency on x.

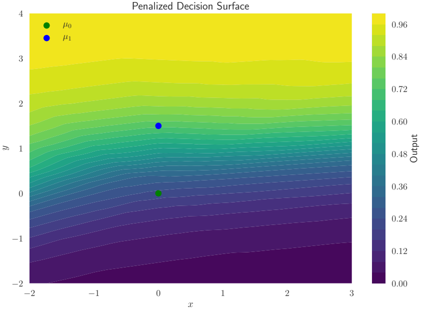


Figure 9: Penalization of objective function straightens contours and reduces x dependency.

Algorithm 1 Encoder Training with Mutual Information-based Regularization

Require: Regularization coefficient $\lambda > 0$, inner learning rate schedule η_t, outer learning rate schedule η'_t

Initialize the parameters of the encoder network E_θ and statistic network f_ϕ.
for $t = 1$ to T **do**
    for $k = 1$ to K **do**
        Sample $\{(x_1, z_1), \ldots, (x_B, z_B)\} \sim \mathbb{P}_{XZ}$ from the joint distribution.
        Sample $\{\tilde{z}_1, \ldots, \tilde{z}_B\} \sim \mathbb{P}_Z$ from the marginal distribution.
        Update f_ϕ by ascending the objective:

$$I_V(\theta, \phi) = \frac{1}{B} \sum_{i=1}^{B} \left[\log \sigma \left(f_\phi(E_\theta(x_i), z_i)\right) - \log\left(1 - \sigma\left(f_\phi(E_\theta(x_i), \tilde{z}_i)\right)\right)\right]$$

$$\phi \leftarrow \phi + \eta_t \nabla_\phi I_V(\theta, \phi)$$

    end for
    Sample $\{(x_1, y_1, z_1), \ldots, (x_B, y_B, z_B)\} \sim \mathbb{P}_{XYZ}$ from the joint distribution.
    Sample $\{\tilde{z}_1, \ldots, \tilde{z}_B\} \sim \mathbb{P}_Z$ from the marginal distribution.
    Update E_θ by descending the objective:

$$\mathcal{L}(\theta, \phi) = \frac{1}{B} \sum_{i=1}^{B} \left[-\log p_\theta\left(y_i | x_i\right) + \lambda I_V\left(\theta, \phi\right)\right]$$

$$\theta \leftarrow \theta - \eta'_t \mathcal{L}(\theta, \phi)$$

end for

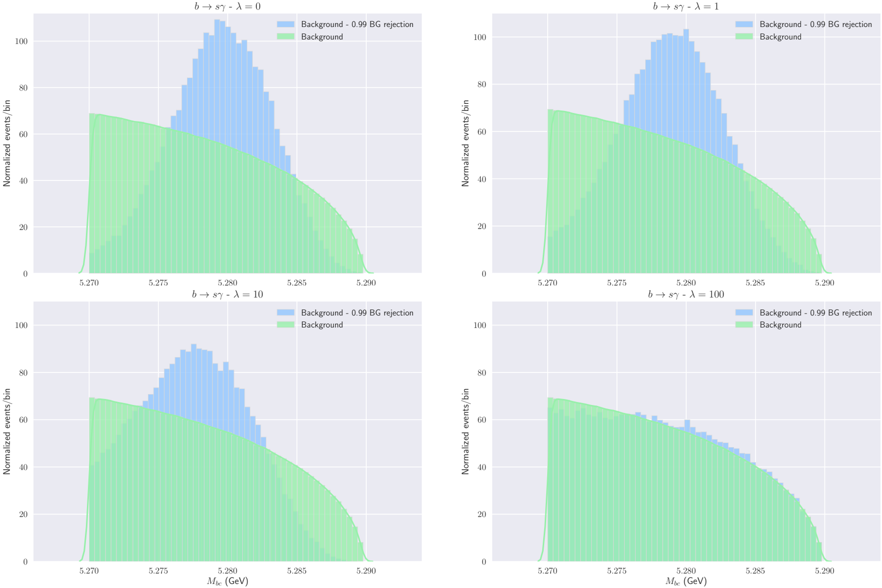
Experiments 2

- Signal classification in FCNC $b \to s\gamma$, want to enforce decorrelation with $Z = M_{bc} \equiv \sqrt{E_{beam}^2 - |\vec{p}_B|^2}$.
- Classifier architecture: 5 layer densely connected network with 512 nodes per layer, SGD ($\eta' = 1\text{e-}4$) w/ Nesterov Momentum ($\gamma = 0.9$).
- Auxillary architecture: 2 dense layers, [256, 128], Adam ($\eta = 1\text{e-}5$)
- Need to use exponential moving average in practice:

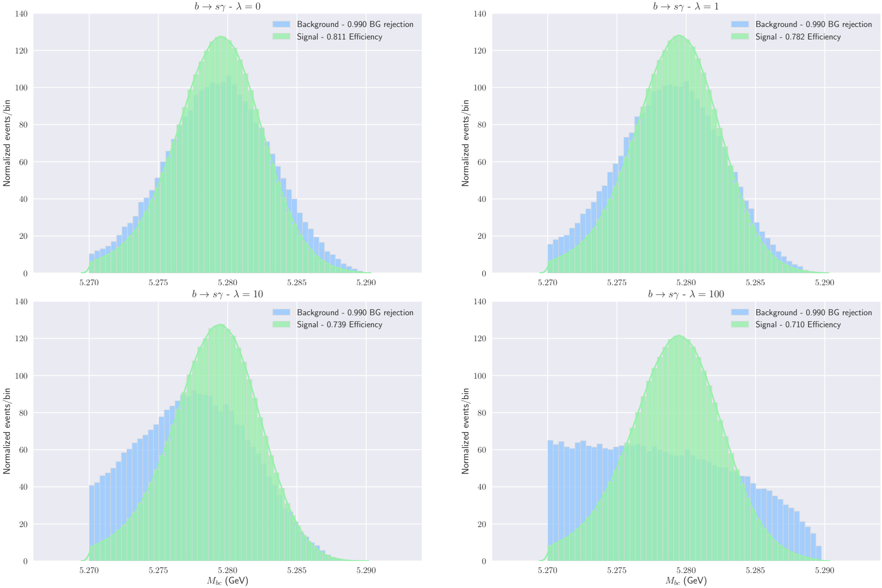
$$\vec{\phi}_{t+1} \leftarrow \vec{\phi}_t + \alpha\left(\vec{\phi}_{t+1} - \vec{\phi}_t\right), \; \alpha \in [0,1]$$

- 7.6M training events, 1.5M test, 5 epochs.

Experiments 2 (Background only)



Experiments 2 (Background + Signal) @ $\epsilon_B = 0.99$



No Free Lunch

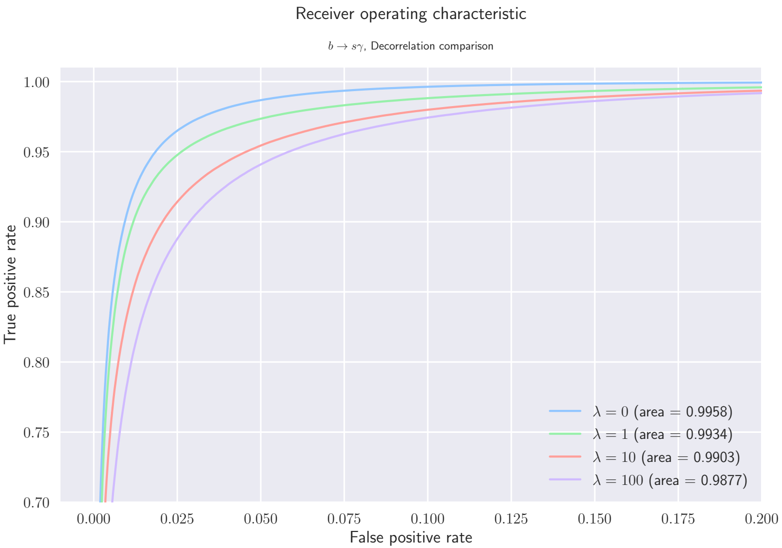


Figure 10: Decorrelation performance penalty. $b \to s\gamma$ events.

No Free Lunch

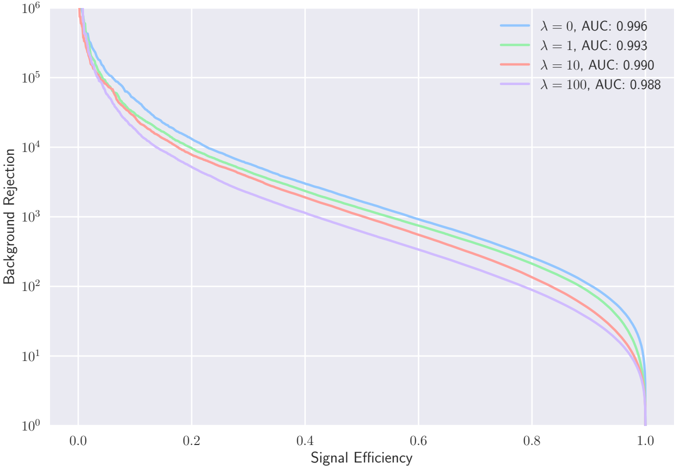


Figure 11: Decorrelation performance penalty. $b \to s\gamma$ events.

Mutual Information as f-Divergence

- f-Divergence: 'Distance' between two probability distributions.

$$d_{KL}(\mu \| \nu) = \mathbf{E}_{x \sim \mu} \left[\log \frac{\mu(x)}{\nu(x)} \right]$$

- Interpretation as KL divergence between joint and product of marginals.

$$\begin{aligned} I(X, Y) &\triangleq \mathbf{E}_{X,Y} \left[\log \frac{p(x,y)}{p(x)p(y)} \right] \\ &= d_{KL} \left(p(x,y) || p(x)p(y) \right) \end{aligned}$$

- Consider the symmetric, bounded form of d_{KL}:

$$\begin{aligned} d_{JS}(\mu || \nu) &\triangleq \frac{1}{2} \left(d_{KL} \left(\mu || m \right) + d_{KL} \left(\nu || m \right) \right) \\ m &= \frac{1}{2}(\mu + \nu) \end{aligned}$$

Jensen-Shannon Divergence Proxy

- Mutual information minimization $\Leftrightarrow$ f-divergence minimization.
- Enforce decorrelation $\Leftrightarrow$ Minimize f-divergence between joint $\mathbb{P}_{XZ}$ and product of marginals $\mathbb{P}_X \otimes \mathbb{P}_Z$.
- Variational lower bound on $d_{JS}\left(\mathbb{P}_X \,\|\, \mathbb{P}_Y\right)$:

$$\begin{aligned} F(\omega) &= \mathbf{E}_{\mathbb{P}_X}\left[\log \sigma\left(T_\omega(x,y)\right)\right] - \mathbf{E}_{\mathbb{P}_Y}\left[\log\left(1-\sigma\left(T_\omega(x,y)\right)\right)\right] \\ &\leq d_{JS}\left(\mathbb{P}_X \,\|\, \mathbb{P}_Y\right) - \log 4 \end{aligned}$$

- Numerically stable. ✔

Mutual Information Penalty v2

- Minimize mutual information between $\mathbb{P}_{XZ}$ and $\mathbb{P}_X \otimes \mathbb{P}_Z$ through minimization of divergence $d_{JS}(\mathbb{P}_{XZ} \,\|\, \mathbb{P}_X \otimes \mathbb{P}_Z)$.

$$
\begin{aligned}
I_V^{(JS)}(E_\theta(X), Z) = \; & \mathbf{E}_{\mathbb{P}_{XZ}} \left[\log \sigma\left(T_\omega(E_\theta(x), z)\right)\right] - \\
& \mathbf{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z} \left[\log\left(1 - \sigma\left(T_\omega(E_\theta(x), z)\right)\right)\right]
\end{aligned}
$$

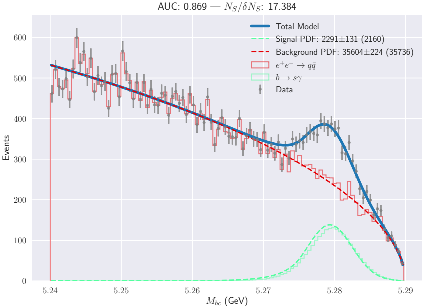
- Numerically stable objective is the sum of two cross-entropy terms. One promotes discrimination power while the other reduces classification dependence on Z.

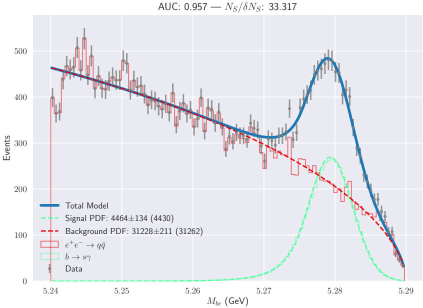
$$
\text{Objective: } \min_E \max_T \mathbf{E}_{\mathcal{D}}\left[-\log p_\theta(y|x)\right] + \lambda_{MI} I_V^{(JS)}\left(E_\theta(X), Z\right) \quad (2)
$$

Comparisons

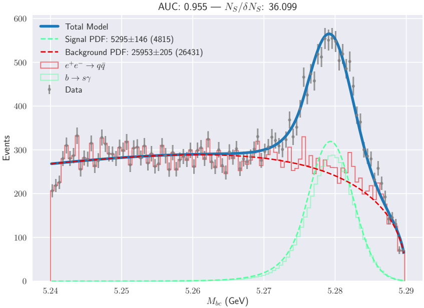
- Use $b \to s\gamma$ sample prepared using centralized Belle II simulation.
  - Investigate effect on fit observables extracted through M_{bc} pdf after 0.999 background rejection.
  - Fix signal PDF shape parameters to original signal sample pre-selection.
  - Float signal/background yields + background shape.
- Optimize for parameter error: $(N_{sig}/\delta N_{sig})$
  - δN_{sig} is the error reported by the covariance matrix.
  - Says nothing about goodness of fit.
  - Metric can probably be improved.

$N_S/\delta N_S$

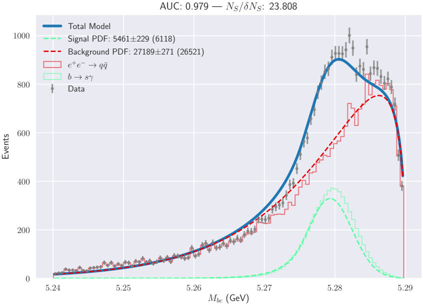

AUC: 0.869 — $N_S/\delta N_S$: 17.384
Total Model
Signal PDF: 2291±131 (2160)
Background PDF: 35604±224 (35736)
$e^+e^- \to q\bar{q}$
$b \to s\gamma$
Data
Events
M_{bc} (GeV)


AUC: 0.957 — $N_S/\delta N_S$: 33.317
Total Model
Signal PDF: 4464±134 (4430)
Background PDF: 31228±211 (31262)
$e^+e^- \to q\bar{q}$
$b \to s\gamma$
Data
Events
M_{bc} (GeV)

$N_S/\delta N_S$



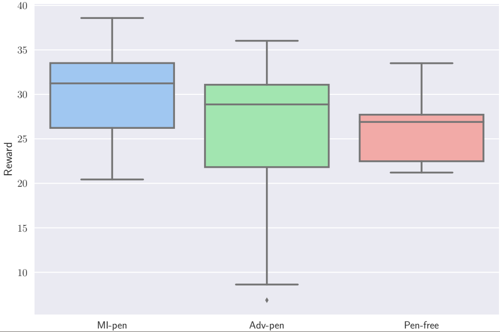
AUC: 0.955 — $N_S/\delta N_S$: 36.099
Total Model
Signal PDF: 5295±146 (4815)
Background PDF: 25953±205 (26431)
$e^+e^- \to q\bar{q}$
$b \to s\gamma$
Data
Events
M_{bc} (GeV)



AUC: 0.979 — $N_S/\delta N_S$: 23.808
Total Model
Signal PDF: 5461±229 (6118)
Background PDF: 27189±271 (26521)
$e^+e^- \to q\bar{q}$
$b \to s\gamma$
Data
Events
M_{bc} (GeV)

Evaluation

- Isolate the effect of fundamental algorithm design from model hyperparameters.
- Run fair automated comparisons with similar techniques. (Fair = same computational budget).
- Select hyperparameters that give high reward $N_S/\delta N_S$.
- Test model sensitivity to hyperparameters by showing distribution of maximum reward achieved by each model (64 samples per model).



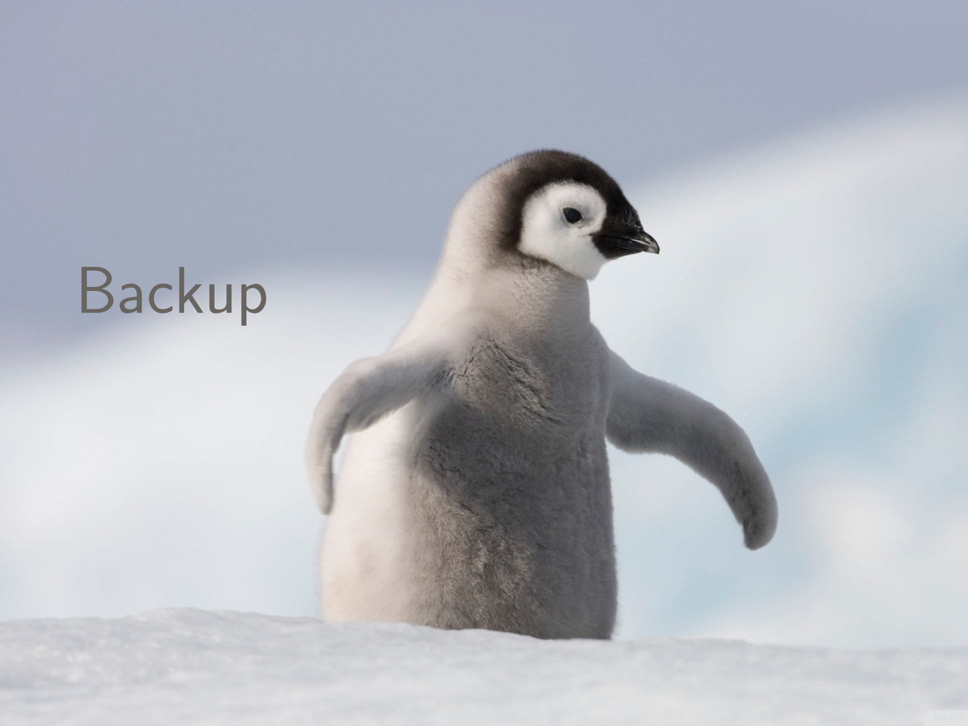
Summary

- Tension between optimal discrimination and systematic errors in searches for NP using ML techniques.
- Methods based on information penalties are an accessible way to prevent background sculpting without significant compromise on discrimination power.

Balance background rejection with controlling systematic uncertainties to achieve better sensitivity to new physics.

justin.tan@coepp.org.au

Backup



Implementation

These gradient-based penalties rely on automatic differentiation frameworks.

- Data collection: `ROOT`
- To Python: `uproot`
- Preprocessing: Spark/Pandas

- Workflow scalable to $\mathcal{O}(100)$ GB worth of training data.
- TensorFlow:
  - Open-source: No black boxes. ✔
  - Fine-grained control over entire architecture. ✔

- Non-SM contributions enter through hypothetical new TeV-scale particles running within the loop $\rightarrow$ interference with known amplitudes.
- Strong constraints on NP by measurement of inclusive/exclusive BR, CP asymmetries.

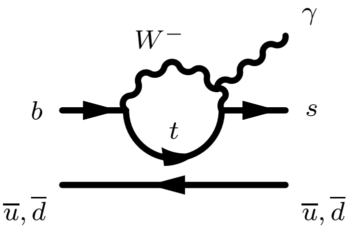


Figure 12: Example of SM radiative penguin decay for $b \rightarrow s\gamma$ [2]

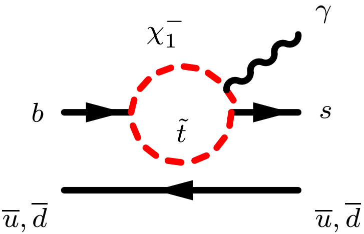


Figure 13: Example of hypothetical SUSY contribution to radiative decay [2]

Mutual Information

- Glossary:
  - $\mathcal{D} = (X, Y)$: True example distribution, $X \in \mathbb{R}^D, y \in [0,1] \sim p$
  - $E = E_{\theta_f}(X) \in \mathbb{R}$: Encoder[1] output parameterized by θ_f, $E \sim q$
  - Z: Variables we would like to remain unbiased
- Want to reduce information content of Z stored in encoding $E_{\theta_f}(X)$.
- Bound $I(E_{\theta_f}(X), Z)$ with Lagrange multiplier λ_{MI}:

$$\mathcal{L}(\theta_f; Z) = H_{p,q} + \lambda_{MI} I(E, Z) \tag{3}$$

- Problem: $I(X, Y)$ between (non-Gaussian) continuous variables intractable.

Experiments 2

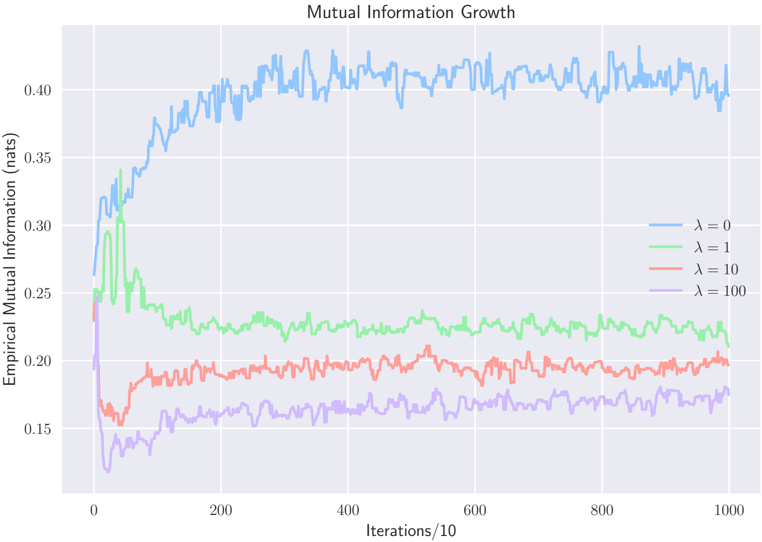


Figure 14: Mutual Information growth over training for different values of λ_{MI}

Algorithm 1 Encoder Training with Mutual Information-based Regularization

Require: Regularization coefficient $\lambda > 0$, inner learning rate schedule η_t, outer learning rate schedule η'_t

Initialize the parameters of the encoder network E_θ and statistic network f_ϕ.

for $t = 1$ to T **do**

  for $k = 1$ to K **do**

    Sample $\{(x_1, z_1), \dots, (x_B, z_B)\} \sim \mathbb{P}_{XZ}$ from the joint distribution.

    Sample $\{\tilde{z}_1, \dots, \tilde{z}_B\} \sim \mathbb{P}_Z$ from the marginal distribution.

    Update f_ϕ by ascending the objective:

$$I_V(\theta, \phi) = \frac{1}{B} \sum_{i=1}^{B} \left[\log \sigma\left(f_\phi(E_\theta(x_i), z_i)\right) - \log\left(1 - \sigma\left(f_\phi(E_\theta(x_i), \tilde{z}_i)\right)\right)\right]$$

$$\phi \leftarrow \phi + \eta_t \nabla_\phi I_V(\theta, \phi)$$

  end for

  Sample $\{(x_1, y_1, z_1), \dots, (x_B, y_B, z_B)\} \sim \mathbb{P}_{XYZ}$ from the joint distribution.

  Sample $\{\tilde{z}_1, \dots, \tilde{z}_B\} \sim \mathbb{P}_Z$ from the marginal distribution.

  Update E_θ by descending the objective:

$$\mathcal{L}(\theta, \phi) = \frac{1}{B} \sum_{i=1}^{B} \left[- \log p_\theta\left(y_i | x_i\right) + \lambda I_V\left(\theta, \phi\right)\right]$$

$$\theta \leftarrow \theta - \eta'_t \mathcal{L}(\theta, \phi)$$

end for

Mutual Information as f-Divergence

- f-Divergence: 'Distance' between two probability distributions.

$$d_{KL}(\mu \| \nu) = \mathbf{E}_{x \sim \mu} \left[\log \frac{\mu(x)}{\nu(x)} \right]$$

- Interpretation as KL divergence between joint and product of marginals.

$$\begin{aligned} I(X, Y) &\triangleq \mathbf{E}_{X,Y} \left[\log \frac{p(x,y)}{p(x)p(y)} \right] \\ &= d_{KL} \left(p(x,y) || p(x)p(y) \right) \end{aligned}$$

- Consider the symmetric, bounded form of d_{KL}:

$$\begin{aligned} d_{JS}(\mu || \nu) &\triangleq \frac{1}{2} \left(d_{KL} \left(\mu || m \right) + d_{KL} \left(\nu || m \right) \right) \\ m &= \frac{1}{2}(\mu + \nu) \end{aligned}$$

Jensen-Shannon Divergence Proxy

- Mutual information minimization $\Leftrightarrow$ f-divergence minimization.
- Variational lower bound on $d_{JS}\left(\mathbb{P}_X \,\|\, \mathbb{P}_Y\right)$:

$$
\begin{aligned}
F(\omega) &= \mathbf{E}_{\mathbb{P}_X}\left[\log \sigma\left(T_\omega(x,y)\right)\right] - \mathbf{E}_{\mathbb{P}_Y}\left[\log\left(1-\sigma\left(T_\omega(x,y)\right)\right)\right] \\
&\leq d_{JS}\left(\mathbb{P}_X \,\|\, \mathbb{P}_Y\right) - \log 4
\end{aligned}
$$

- Numerically stable. ✔
- Enforce decorrelation $\Leftrightarrow$ Minimize f-divergence between joint $\mathbb{P}_{XZ}$ and product of marginals $\mathbb{P}_X \otimes \mathbb{P}_Z$.

Mutual Information Penalty v2

- Minimize mutual information between $\mathbb{P}_{XZ}$ and $\mathbb{P}_X \otimes \mathbb{P}_Z$ through minimization of divergence $d_{JS}(\mathbb{P}_{XZ} \,\|\, \mathbb{P}_X \otimes \mathbb{P}_Z)$.

$$
\begin{aligned}
I_V^{(JS)}(E_\theta(X), Z) = \mathbf{E}_{\mathbb{P}_{XZ}}\left[\log \sigma\left(T_\omega(E_\theta(x), z)\right)\right] - \\
\mathbf{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}\left[\log\left(1 - \sigma\left(T_\omega(E_\theta(x), z)\right)\right)\right]
\end{aligned}
$$

- Numerically stable objective is the sum of two cross-entropy terms. One promotes discrimination power while the other reduces classification dependence on Z.

$$
\text{Objective: } \min_E \max_T \mathbf{E}_{\mathcal{D}}\left[-\log p_\theta(y|x)\right] + \lambda_{MI} I_V^{(JS)}\left(E_\theta(X), Z\right) \quad (4)
$$

f-Divergences

- Measure disimilarity between two given probability distributions.

$$d_f(\mu \| \nu) \triangleq \int_{\mathcal{X}} \nu(x) f\left(\frac{\mu(x)}{\nu(x)}\right)$$

  - Generator $f : \mathbb{R}^+ \to \mathbb{R}$ convex with $f(1) = 0$
  - KL-Divergence: $f(v) = v \log v$

- Variational lower bound by applying Jensen's inequality to Fenchel dual-dual.

$$d_f(\mu \| \nu) \geq \sup_{T \in \mathcal{T}} \left(\mathbf{E}_{x\sim\mu}\left[T(x)\right] - \mathbf{E}_{x\sim\nu}\left[f^*(T(x))\right]\right)$$

  - $\mathcal{T}$: Arbitrary class of functions $T : \mathcal{X} \to \mathbb{R}$
  - f^*: Fenchel dual $f^*(t) \triangleq \sup_{u \in \mathrm{Dom}_f} (ut - f(u))$
  - Estimated using Monte Carlo sampling.