テキスト生成における最小ベイズリスク復号の理論的な理解に向けて

市原有生希 [1]　陣内佑 [2]　蟻生開人 [2]　森村哲郎 [2]　内部英治 [3]
[1] 奈良先端科学技術大学院大学　[2] サイバーエージェント
[3] 国際電気通信基礎技術研究所
ichihara.yuki.iu1@is.naist.jp
{jinnai_yu,kaito_ariu,morimura_tetsuro}@cyberagent.co.jp
uchibe@atr.jp

概要

最小ベイズリスク復号 (Minimum Bayes Risk decoding) は, 自然言語処理のテキスト生成において効果的であることが知られている手法である. この手法は基盤となる人間の嗜好確率分布に基づく期待効用を最大化することを目的とし, 出力選択を行う. 先行研究における実験的評価ではこのアプローチが顕著な成功を収めていることが示されているが, これらの手法が有効に機能する原因については未だ解明されていない. 本研究では最小ベイズリスク復号が何故高い性能が得られるかを明らかにすることを目的として, その理論的な性能を分析する. 分析の結果として, いくつかの仮定の下, 最小ベイズリスク復号の誤差が計算に用いる参照仮説集合の大きさ n に対して高い確率で $O(\frac{1}{\sqrt{n}})$ に収まることが示された.

1　はじめに

最小ベイズリスク (Minimum Bayes Risk) 復号 [1, 2] は, 自己回帰型確率モデル (例: 大規模言語モデル) から系列を生成するための決定則である. 最小ベイズリスク復号は, 機械翻訳 [3, 4], 画像キャプション生成 [5], および指示追従型タスク [6] など, さまざまなテキスト生成タスクにおいて高品質なテキストを生成できることが示されている. また, 多くの実験で, 最小ベイズリスク復号が MAP 復号 (例: ビームサーチ)[7, 8] に比べて優位性を持つことが報告されている.

最小ベイズリスク復号の有効性は, 多くの実験的評価によって示されているが, その根本的な理由については十分に解明されていない. Kamigaito ら [9] は最小ベイズリスク復号の人間とモデルの嗜好分布間の不一致はバイアスと多様性の項に分解できることを示唆した. Bertsch ら [10] は最小ベイズリスク復号の性能の成功に寄与する 4 つの要因を実験的に分析した. さらに, Ohashi ら [11] は異常検知を用いた解析によって, モデルの嗜好分布が人間の嗜好分布に近い場合, 最小ベイズリスク復号が優れた性能を引き起こすことを確認した.

本研究では, 最小ベイズリスク復号による出力の理論的な性能の解析を行う. 本研究の結果は略式に述べると以下になる.

定理 1 (最小ベイズリスク復号の収束レート; 略式)**.** 後述する仮定の下, 最小ベイズリスク復号の誤差は高い確率で参照仮説集合の大きさ n に対して $O\left(\frac{1}{\sqrt{n}}\right)$ である.

この理論的な知見は, 先行研究において最小ベイズリスク復号の性能が参照仮説集合の大きさ n に対して向上するという実験的評価と整合している [8, 12]. 最小ベイズリスク復号によって生成されるテキストの品質に関する実験的評価は多く先行研究があるが, 理論的な誤差の上界を解析した研究はこれまで存在していない. 本結果はテキスト生成アルゴリズムにおける未解決の課題の一つに対する解答につながるものである.

2　最小ベイズリスク復号

テキスト生成は, 入力列 x に対して出力列 y を生成するタスクである. テキスト生成モデルは, 仮説の出力空間 Y 上に確率分布 $P_{\text{model}}(y \mid x)$ を定義する. 完全な仮説の集合 Y は次のように定義される:

$$\mathrm{Y} := \{\mathrm{BOS} \circ \mathbf{v} \circ \mathrm{EOS} \mid \mathbf{v} \in \mathrm{V}^*\}.$$

ここで, $\circ$ は文字列の連結を表し, V^* は語彙集合 V の Kleene 閉包である. 復号は, 与えられた入力に対して最もスコアの高い仮説を見つけることを目的とする.

最小ベイズリスク復号は, 期待効用を最大化する出力列を求める枠組みであり [1, 2], テキスト生成モデル P_{model} と効用関数 $u(h, y)$ の 2 つの要素から構成される. ここで, $u(h, y)$ は候補出力 h の品質を, 参照出力 y をもとに定量化する指標である. 記号の簡略化として, 条件付き分布 $P(h \mid x)$ を $P(h)$ と略記する.

理想的な状況では, モデルではなく人間の嗜好分布 P_{human} に基づく期待効用最大化問題を考えたい. すなわち, 参照仮説集合 Y を用いて, 以下のような最適な出力 h^{human} を定義する:

$$u_h(h) := \sum_{y \in \mathrm{Y}} u(h, y) \cdot P_{\text{human}}(y). \tag{1}$$

$$h^{\text{human}} := \arg\max_{h \in \mathrm{H}} u_h(h). \tag{2}$$

ここで, P_{human} は人間の嗜好分布, H は候補仮説集合, $u : \mathrm{H} \times \mathrm{Y} \to [0, U]$ は効用関数である. 加えて, 本研究は $\mathrm{H} = \mathrm{Y}$ が成り立つことを仮定する. 実際には, P_{human} を直接利用することは不可能であるため, 最小ベイズリスク復号では, モデルの嗜好分布 P_{model} を人間の嗜好分布 P_{human} の近似として用いる. その結果, 以下のような最適な出力 h^{model} を求める問題が得られる.

$$u_m(h) := \sum_{y \in \mathrm{Y}} u(h, y) \cdot P_{\text{model}}(y). \tag{3}$$

$$h^{\text{model}} := \arg\max_{h \in \mathrm{H}} u_m(h). \tag{4}$$

しかし, Y 全体にわたる期待値の計算は, 実行困難であるため, 実用上はモデル P_{model} からサンプリングした参照仮説集合 $\mathrm{Y}^{\mathrm{n}}_{\text{ref}} = \{y_1, \ldots, y_n\}$ を用いて, この期待値をモンテカルロ法で近似する [8, 13]. これにより, 有限個のサンプルを参照とした問題となる:

$$\widehat{u}(h) := \frac{1}{n} \sum_{y \in \mathrm{Y}^{\mathrm{n}}_{\text{ref}}} u(h, y) \tag{5}$$

$$h^{\text{mc}} := \arg\max_{h \in \mathrm{H}_{\text{ref}}} \widehat{u}(h). \tag{6}$$

ここで, H_{ref} は $\mathrm{Y}^{\mathrm{n}}_{\text{ref}}$ を用いて構成される候補仮説集合とし, n をその大きさ $n = |\mathrm{Y}^{\mathrm{n}}_{\text{ref}}|$ とする. $n \to \infty$ のとき, $\mathrm{Y}^{\mathrm{n}}_{\text{ref}}$ が全集合 Y と一致することが期待でき, この近似は, 理論的に完全な最小ベイズリスク復号の目的へと帰着する.

以上のように, 実用上では, 最小ベイズリスク復号は, モデルの嗜好分布を用いて, 間接的に人間の嗜好分布を近似し, 効用最大化の視点から, 最も良いと評価される候補を選択する枠組みである.

3 理論解析

本研究では, $u_h(h^{\text{human}})$ を目標として議論を進めていくが, その理由を以下に述べる. 直感的には, 以下の式のように, 効用関数を用いずにモデルの嗜好分布のみを基にして, 最頻値を選択する手法 (MAP 復号) が最適であると考えられる.

$$\textbf{MAP 復号の目的式} := \arg\max_{h \in \mathrm{H}} P_{\text{model}}(h). \tag{7}$$

しかし, 出力確率の最頻値を考慮する場合 (式 7), 人間らしくない出力を引き起こす問題が指摘されている [14]. このような課題にも対処可能であり, 実験評価においても MAP 復号と比較して高性能を示し, さらに Workshop on Statistical Machine Translation (WMT) ではタスク評価でも使用されていることが多いことから, より優れた評価指標と考えられる. モデルの嗜好分布が人間の嗜好分布に近い場合に性能が良くなる結果 [11] から, 最小ベイズリスク復号の目的 h^{human} を私たちが求めるべき真の値であると仮定して, 解析を進めていく.

3.1 事前知識

本実験の解析で, 用いた 2 つの集中不等式を以下に示す.

定理 2. 一様集中不等式 (Theorem 4.10 [15])
$\mathcal{F}$ を関数の集合, $f \in \mathcal{F} : X_i \to [0, b]$ とする.

$$\Pr\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \le 2\mathcal{R}_n(\mathcal{F}) + \epsilon\right) \le 1 - \exp\left(-\frac{n\epsilon^2}{2b^2}\right).$$

ここで, $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{P} f|$, $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$, $\mathbb{P} f = \mathbb{E}[f(X)]$ であり, X および $\{X_i\}_{i=1}^n$ は $\mathbb{P}$ から独立同分布でサンプリングされたものとする. また, $\mathcal{R}_n : \left(\mathcal{F}, \{X_i\}_{i=1}^n\right) \to \mathbb{R}$ をラデマッハ複雑度 (関数がどれだけ複雑かを示す指標) とする.

次に, Hoeffding の不等式を以下に示す.

定理 3. Hoeffding の不等式 (Corollary 1.1 [16])
$\{X_i\}_{i=1}^n \in [0, b]$ とし, 独立同分布でサンプリングされたものとする.

$$\Pr\left(\left|\mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^n X_i\right| \le \epsilon\right) \le 1 - 2\exp\left(-\frac{2n\epsilon^2}{b^2}\right).$$

3.2 問題設定

ここで, 人間の嗜好分布 P_{human} と効用関数 u は, 人間の好みを反映した分布および関数であると仮定す

る. このとき, P_{human} 下での最適な出力は h^{human} である. しかしながら, 実用上の制約から, 実際にはモンテカルロ法による擬似最適解 h^{mc} しか得られない. 一方で, この h^{mc} は最終的に人間の嗜好分布 P_{human} 下で評価されるため, 以下のような性能差が生じる.

実用的に使われている最小ベイズリスク復号と, 真の最小ベイズリスク復号の最適解を, 人間の嗜好分布下で評価した場合の誤差を以下に示す.

$$\text{Regret}(h^{\text{human}}, h^{\text{mc}}) := u_h(h^{\text{human}}) - u_h(h^{\text{mc}}). \quad (8)$$

本研究の目標は, この性能差の上界 (式 (8)) を理論的に求めることにある. もし, この上界をサンプル数に関するオーダーで示すことができれば, モンテカルロ法を用いた最小ベイズリスク復号の性能に対して, 理論的な保証を提供できることになる.

ここで, 本研究では, 以下の仮定が成り立つものとし, 解析を進める.

仮定 1. P_{model} は, P_{human} から得られた $|D|$ 個のサンプルによる経験分布とする.

このとき, $P_{\text{model}}, u_m(h)$ は以下のように表される.

$$P_{\text{model}}(y) = \frac{1}{|D|} \sum_{y' \in D} I(y = y'). \quad D \sim P_{\text{human}}(\cdot \mid x)$$

ここで, I は指示関数とする.

$$u_m(h) = \frac{1}{|D|} \sum_{y \in D} u(h, y).$$

仮定 1 の解釈の例として, 一般に, P_{model} は大規模言語モデルが想定されるが, 本研究では理論解析のため, P_{model} に対してこのようなサンプルベースの分布を仮定している. 大規模言語モデルが大量の学習データから構築されることを考慮すると, $|D| \gg n$ という条件が自然に成立すると考えられる

仮定 2. 本研究の設定に基づき, 全ての y および h に対して, 効用関数 u が線形関数として表現できるような 埋め込み関数 $\alpha(h) \in \mathbb{R}^d$ および $\mathbf{v}(y) \in \mathbb{R}^d$ が存在すると仮定する. さらに, $\mathbf{v}(y)$ の各要素は部分空間の正規直交基底であると想定する.

$$u(h, y) = \alpha(h)^\top \mathbf{v}(y).$$

このような性質を満たす埋め込み関数としては, 文の類似度を評価するための Sentence Bert や Sentence Transformer などの例が存在する [17, 18, 19].

3.3 理論解析

式 (8) に基づき, モンテカルロ法によって得られた h^{mc} の性能が, 真の最小ベイズリスク復号の値から, どの程度乖離しているかを解析する.

3.3.1 $\text{Regret}(h^{\text{human}}, h^{\text{mc}})$ の上界

$\text{Regret}(h^{\text{human}}, h^{\text{mc}})$ の上界について以下の定理が成り立つ.

定理 1. 仮定 1, 仮定 2 の元で, 次の上界が, 少なくとも確率 $1 - \delta$ で成立する:

$$\text{Regret}(h^{\text{human}}, h^{\text{mc}}) \le 2U\sqrt{\frac{1}{2n}\log\frac{8}{\delta}} + \frac{12U}{n}\left(\sqrt{d\log(2\sqrt{d})} + 2\sqrt{d}\right) + 2U\sqrt{\frac{1}{2|D|}\log\frac{8}{\delta}}.$$

この上界は, $O\left(\max\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{|D|}}\right)\right)$ の速さで減衰していくことがわかる.

この定理は, 有限サンプルから推定した最小ベイズリスク復号に対して, 真の最小ベイズリスク復号との性能差がどの程度まで抑えられるかを定量的に示す重要な結果である. 本結果は, 参照仮説集合の大きさ n および $|D|$ の増大に伴って性能差が縮小することを保証する.

ここから, 定理 1 の証明を行う. 式 (8) は分解すると, 以下の 3 つの式に分けることができる.

$$\text{Regret}(h^{\text{human}}, h^{\text{mc}}) = \underbrace{u_h(h^{\text{human}}) - u_m(h^{\text{human}})}_{\clubsuit} + \underbrace{u_m(h^{\text{human}}) - u_m(h^{\text{mc}})}_{\heartsuit} + \underbrace{u_m(h^{\text{mc}}) - u_h(h^{\text{mc}})}_{\spadesuit}.$$

クローバーおよびスペード項の上界 まず, クローバー $\clubsuit$ およびスペード $\spadesuit$ 項を解析する. ここで, $\clubsuit$ 項は人間の嗜好分布 P_{human} とモデルの嗜好分布 P_{model} における 真の最適出力 h^{human} の性能差を表し, $\spadesuit$ 項は近似最適出力 h^{mc} に対し, 人間の嗜好分布とモデルの嗜好分布間での性能差を示している. 仮定 1 に基づき, P_{model} は P_{human} のサンプルを用いて構成されているため, Hoeffding の不等式 (定理 3) を適用可能であり, 以下の補題が成り立つ.

補題 1. 仮定 1 の元で, 次の上界が, 少なくとも確率 $1 - \frac{\delta}{2}$ で成立する:

$$\clubsuit + \spadesuit \le 2U\sqrt{\frac{1}{2|D|}\log\frac{8}{\delta}}.$$

証明 仮定 1 の元で, 定理 3 を適用すると,$\clubsuit$ 項は以下のように表される:

$$\Pr\left(|\clubsuit| \le \epsilon\right) \le 1 - 2\exp\left(-\frac{2|D|\epsilon^2}{U^2}\right) = 1 - \frac{\delta}{4}.$$

ここで δ は任意の正の値とする. ここで, ϵ について

整理すると, 次のように表すことができる.

$$\epsilon = U\sqrt{\frac{1}{2|D|}\log\left(\frac{8}{\delta}\right)}.$$

この ϵ は, サンプル数 $|D|$ が分母に含まれる形で記述されており, $|D|$ を増やせば, 0 に収束することは直感的に理解できる. 今回の場合において, Hoeffding の不等式が意味していることは, ♣ 項が ϵ よりも小さくなる確率が, 少なくとも $1-\frac{\delta}{4}$ であることを示唆している. すなわち, 以下のような ♣ の上界は少なくとも確率 $1-\frac{\delta}{4}$ で成立する:

$$\clubsuit \leq U\sqrt{\frac{1}{2n}\log\left(\frac{8}{\delta}\right)}.$$

♠ 項に関しても, 同様の操作によって上界を求めることができ, 補題 1 が証明される.

□

これらのことから, ♣ 項と ♠ 項の上界は, サンプル数 $|D|$ のみが変化し, それの大きさに応じて, 上界が $\frac{1}{\sqrt{|D|}}$ の速度で減衰していくことを示唆している.

ハート項の上界　次に, ハート ♡ 項の解析から始める. ♡ 項はモデルの嗜好分布下における最適出力 h^{human} と近似最適出力 h^{mc} の性能差を示唆している. ♡ を次のように分解する.

$$\begin{aligned} &u_m(h^{\text{human}}) - u_m(h^{\text{mc}}) = u_m(h^{\text{human}}) - \widehat{u}(h^{\text{human}}) \\ &+ \widehat{u}(h^{\text{mc}}) - u_m(h^{\text{mc}}) + \underbrace{\widehat{u}(h^{\text{human}}) - \widehat{u}(h^{\text{mc}})}_{\leq 0} \\ &\leq \underbrace{u_m(h^{\text{human}}) - \widehat{u}(h^{\text{human}})}_{\heartsuit_1} + \underbrace{\widehat{u}(h^{\text{mc}}) - u_m(h^{\text{mc}})}_{\heartsuit_2}. \end{aligned}$$

これらの定理 2 , 定理 3 および仮定 2 を用いて, $\heartsuit_1$ と $\heartsuit_2$ の上界を順に示していく.

補題 2. 仮定 2 の元で, 次の上界が, 少なくとも確率 $1-\frac{\delta}{2}$ で成立する:

$$\heartsuit \leq 2U\sqrt{\frac{1}{2n}\log\frac{8}{\delta}} + \frac{12U}{n}\left(\sqrt{d\log(2\sqrt{d})} + 2\sqrt{d}\right).$$

証明　$\heartsuit_1$ 項は, 補題 1 の証明と同様の操作によって求めることができ, 以下のような上界が少なくとも確率 $1-\frac{\delta}{4}$ で成立する:

$$\heartsuit_1 \leq U\sqrt{\frac{1}{2n}\log\left(\frac{8}{\delta}\right)}.$$

次に $\heartsuit_2$ 項 を解析する. ここで, $\heartsuit_1$ 項の解析と同様に, Hoeffding の不等式を適用することはできない. これは, h^{mc} を選択する際に $\widehat{u}$ に依存しているためである. この理由から, 定理 2 を適用して解析を行う.

$\heartsuit_2$ 項の上界は少なくとも確率 $1-\frac{\delta}{4}$ で成立する:

$$\heartsuit_2 \leq U\sqrt{\frac{1}{2n}\log\left(\frac{4}{\delta}\right)} + 2\mathcal{R}_n(\mathcal{F}).$$

[20] の 27.2 章より, ラデマッハ複雑度 $\mathcal{R}_n(\mathcal{F})$ に関する次の上界が得られる:

$$2\mathcal{R}_n(\mathcal{F}) \leq \frac{12U}{n}\left(\sqrt{d\log(2\sqrt{d})} + 2\sqrt{d}\right).$$

□

これらの結果から, $\heartsuit_1$ と $\heartsuit_2$ 項の上界は, サンプル数 n のみが変化し, それの大きさに応じて, 上界が $\frac{1}{\sqrt{n}}$ の速度で減衰していくことを示唆している.

これらの補題 1, 補題 2 を用いることで, 最終的に $\text{Regret}(h^{\text{human}}, h^{\text{mc}})$ の上界 (定理 1) が導出される.

4　おわりに

本研究では, 最小ベイズリスク復号に関する理論的性能保証を初めて示し, 最小ベイズリスク復号が, 高品質な出力を得るための有効な手法であることを数理的観点から示唆した. 具体的には, モデルの嗜好分布に基づくモンテカルロ法を用いた最小ベイズリスク復号において, 真の解との性能差の上界を導出し, その性能差が参照仮説集合の大きさ n の増大とともに $O(\frac{1}{\sqrt{n}})$ オーダーで収束することを示した. 総じて, 本研究は最小ベイズリスク復号の数理的理解を深化させた.

謝辞

本研究は JSPS 科研費 23K19986 の助成を受けたものです.

参考文献

[1] Shankar Kumar and William Byrne. Minimum Bayes-Risk Word Alignments of Bilingual Texts. In **Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)**, pp. 140–147. Association for Computational Linguistics, July 2002.

[2] Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In **Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004**, pp. 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[3] Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. Minimum Bayes Risk Combination of Translation Hypotheses from Alternative Morphological Decompositions. In Mari Ostendorf, Michael Collins, Shri Narayanan, Douglas W.

Oard, and Lucy Vanderwende, editors, **Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers**, pp. 73–76, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[4] Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. Neural Machine Translation by Minimising the Bayes-risk with Respect to Syntactic Translation Lattices. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 362–368, Valencia, Spain, April 2017. Association for Computational Linguistics.

[5] Sebastian Borgeaud and Guy Emerson. Leveraging Sentence Similarity in Natural Language Generation: Improving Beam Search using Range Voting. In Alexandra Birch, Andrew Finch, Hiroaki Hayashi, Kenneth Heafield, Marcin Junczys-Dowmunt, Ioannis Konstas, Xian Li, Graham Neubig, and Yusuke Oda, editors, **Proceedings of the Fourth Workshop on Neural Generation and Translation**, pp. 97–109, Online, July 2020. Association for Computational Linguistics.

[6] Ian Wu, Patrick Fernandes, Amanda Bertsch, Seungone Kim, Sina Pakazad, and Graham Neubig. Better instruction-following through minimum bayes risk. **arXiv preprint arXiv:2410.02902**, 2024.

[7] Mathias Müller and Rico Sennrich. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 259–272, Online, August 2021. Association for Computational Linguistics.

[8] Bryan Eikema and Wilker Aziz. Sampling-Based Approximations to Minimum Bayes Risk Decoding for Neural Machine Translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 10978–10993, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[9] Hidetaka Kamigaito, Hiroyuki Deguchi, Yusuke Sakai, Katsuhiko Hayashi, and Taro Watanabe. Theoretical Aspects of Bias and Diversity in Minimum Bayes Risk Decoding. **arXiv preprint arXiv:2410.15021**, 2024.

[10] Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew Gormley. It's MBR all the way down: Modern generation techniques through the lens of minimum Bayes risk. In Yanai Elazar, Allyson Ettinger, Nora Kassner, Sebastian Ruder, and Noah A. Smith, editors, **Proceedings of the Big Picture Workshop**, pp. 108–122, Singapore, December 2023. Association for Computational Linguistics.

[11] Atsumoto Ohashi, Ukyo Honda, Tetsuro Morimura, and Yuu Jinnai. On the True Distribution Approximation of Minimum Bayes-Risk Decoding. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)**, pp. 459–468, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[12] Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 811–825, 2022.

[13] António Farinhas, José de Souza, and Andre Martins. An Empirical Study of Translation Hypothesis Ensembling with Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 11956–11970, Singapore, December 2023. Association for Computational Linguistics.

[14] Clara Meister, Gian Wiher, Tiago Pimentel, and Ryan Cotterell. On the probability–quality paradox in language generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 36–45, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[15] Martin J Wainwright. **High-dimensional statistics: A non-asymptotic viewpoint**, Vol. 48. Cambridge university press, 2019.

[16] Francis Bach. **Learning Theory from First Principles**. Adaptive Computation and Machine Learning series. MIT Press, 2024.

[17] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[18] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4512–4525, Online, November 2020. Association for Computational Linguistics.

[19] Shafiq Rayhan Joty Caiming Xiong Yingbo Zhou Semih Yavuz Rui Meng*, Ye Liu*. Sfr-embedding-2: Advanced text embedding with multi-stage training, 2024.

[20] Shai Shalev-Shwartz and Shai Ben-David. **Understanding machine learning: From theory to algorithms**. Cambridge university press, 2014.